Calculate Z statistic

n_{ILEC} = 145,629 installation orders $\sigma_{ILEC} = 1.40621$

n_{CLEC} = 571 installation orders

STEP #2:

$$\sigma_{DIFF} = \sigma_{ILEC} * \text{sqrt}[1/ n_{ILEC} + 1/n_{CLEC}]$$
$$= 1.40621 * \text{sqrt}[1/145{,}629 + 1/571] = 0.05896$$

STEP #3:

$$\mathbf{Z = DIFF/ \sigma_{DIFF} = 0.57228 / 0.05896 = 9.71}$$

Compare to Critical Z

Z = 9.71

Since there were 38 submeasures for XYZ in Illinois reported this month*, Critical Z = 1.68

Since 9.71 > 1.68,

Test fails!!

Number of Performance Tests	K Values	Critical Z-value
1	0	1.65
2	0	1.96
3	0	2.12
4	0	2.23
5	0	2.32
6	0	2.39
7	0	2.44
8	1	1.69
9	1	1.74
10-19	1	1.79
20-29	2	1.73
30-39	**3**	**1.68**
40-49	3	1.81
50-59	4	1.75
60-69	5	1.70

* Includes only measures eligible for remedies with at least 10 data points.

C

ATTACHMENT C

Before the
FEDERAL COMMUNICATIONS COMMISSION
Washington, D.C. 20554

In the Matter of)
)
Performance Measurements and)
Reporting Requirements) CC Docket No. 98-56
for Operations Support Systems,) RM 9101
Interconnection, and Operator)
Services and Directory)
Assistance)

Affidavit of Dr. Colin L. Mallows

Colin L. Mallows, being duly sworn, deposes and says:

1. I am a Technology Consultant at AT&T Laboratories. I make this affidavit in support of AT&T's comments regarding the use of statistical methods to determine whether incumbent local exchange carriers ("ILECs") are providing nondiscriminatory, i.e., parity, service to competing carriers ("CLECs"). I understand this is a requirement of law under Section 251 of the Telecommunications Act of 1996 ("Act").

Qualifications

2. I have been a professional statistician for nearly 45 years. I obtained a B.Sc. in Mathematics in 1951 and a Ph.D. in Statistics in 1953, both from University College, London. After two years in the British Army I became a lecturer at University College in the area of statistics. Since 1960, I have been employed at AT&T (formerly Bell)

Laboratories, becoming Head of the Statistical Models and Methods Research Department in 1969. I relinquished that title in 1986. From 1960 through 1964, I was also an adjunct associate professor at Columbia University, teaching courses in statistical analysis.

3. I am a Fellow of the American Statistical Association ("ASA"), and I served as an associate editor of Journal of the American Statistical Association from 1966 to 1971, and again from 1986-1989. I am also a Fellow of the Institute of Mathematical Statistics ("IMS"), and an elected member of the International Statistical Institute. I was twice elected to the Council of IMS, and have served on various committees of the IMS and ASA. In 1997 I was honored by being named Fisher Lecturer at the Joint Statistical Meetings held by the ASA, IMS, the International Biometric Society and the Statistical Society of Canada.

4. I have published over 100 papers, with a large number of co-authors, in a variety of journals. My name is attached to several well-known statistical techniques, including the Cp-plot for selecting regression variables, the phi-model for analysis of ranking data, and a weighting scheme for robust linear regression. My professional interests include foundations, data analysis, statistical graphics, time series, robustness, software reliability,

moment-problems and Chebychev inequalities, combinatorics and coding theory.

Introduction

5. I have reviewed the Commission's Notice of Proposed Rulemaking ("Notice") in this proceeding, focusing on its discussion of the use of statistical analysis as a means of determining whether ILECs are providing parity service to new competitors. The Notice (¶ 34) is clearly correct that "reporting averages of performance measurements alone, without further analysis, may not reveal whether there are underlying differences in the way incumbent LECs treat their own retail operations in relation to the way they treat competing carriers." Thus, it properly proposes to require the use of statistical tests to determine whether measured differences in average ILEC performance for themselves and competitors "represent true differences in behavior rather than random chance."

6. As the Commission is aware, AT&T has supported the use of statistical tests to determine whether an ILEC has met its statutory obligations. Earlier this year, AT&T provided the Commission with a concept for applying statistical analysis to ILEC performance measurements.[1] The AT&T Statistical Ex Parte provided a methodology, given the

[1] _Ex parte_ letter from Frank S. Simone, AT&T to Magalie Roman Salas, FCC, CC Docket No. 96-98, RM9101, dated February 3, 1998 ("AT&T Statistical Ex Parte").

presence of random error, to determine if an ILEC has complied with its statutory obligations when it reports results of numerous individual parity measurements, some of which show "worse" results for CLECs than for the ILEC.[2]

7. AT&T's Statistical Ex Parte correctly recognized that each of the individual tests of ILEC performance contained statistical Type I error. Thus, it is appropriate to use a Type I error concept when reviewing the ILEC's parity tests in the aggregate to determine whether the ILEC has met its nondiscrimination obligations. AT&T's Statistical Ex Parte thus described the use of a three-part analysis to determine whether ILEC measurements and reported results, when viewed in the aggregate, represent nondiscriminatory performance.[3]

8. Since that time, I have been asked to review and comment upon AT&T's Statistical Ex Parte and provide additional insight on the use of statistical tests in this

[2] Since most of the measurements for these purposes are measurements of time, a "worse" result for a CLEC is usually a larger value, e.g., a 5-day installation interval for a CLEC is worse than a 3-day interval for the ILEC.

[3] AT&T's proposal recommended establishment of separate thresholds for: (1) the maximum number of "failures" on a monthly report that could reasonably represent mere randomness resulting from the measurement process rather than disparity of performance; (2) repeated failures on specific performance measurements in consecutive months; and (3) measurements showing extreme differences in average performance for the ILEC and CLECs. Id., p. 3.

context. As described in Section I below, the more detailed statistical methodology that is proposed here requires only a two-part analysis and provides the ILECs with more leeway than the original AT&T proposal. Nevertheless, I believe that it provides a valid statistical comparison of the ILECs' actual performance for itself and CLECs.

Summary of Testimony

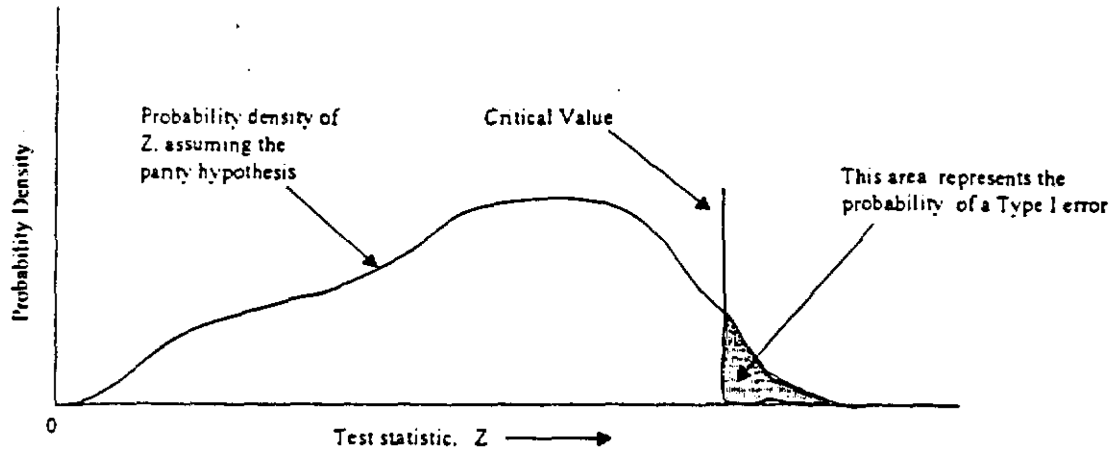
9. Specifically, my testimony below shows that AT&T's proposed methodology satisfies the Commission's desire to assure that reported differences in ILEC performance are statistically meaningful. With respect to individual tests of ILEC performance, there are three key components in developing an appropriate statistical methodology. First, the modified z-statistic proposed by LCUG provides an appropriate test statistic to determine whether there are significant differences in the mean and the variance of an ILEC's performance for itself and for CLECs. Second, a one-tailed test with Type I error held at the 5% level strikes a fair balance between the need to account for both Type I and Type II errors. Third, the t-distribution provides a useful basis for calculating the critical value for individual tests of ILEC performance, which is used to determine whether CLECs have been given equal treatment by the ILEC. Moreover, in those cases where the sizes of the ILEC and

measurements, from which the observed measurements are assumed to be drawn. We cannot observe these populations, and must base our test procedures on the observed samples. If the null hypothesis is accepted through the use of the chosen tests, then any differences in the ILEC's performance results for itself and the CLEC are deemed "statistically insignificant," and parity can be assumed.

13. All such statistical tests have three components. First, the test designer must select a test statistic, which is a formula that produces a single number summarizing the observed ILEC and CLEC data. Next, an acceptable Type I error probability must be adopted. The error probability represents the test designer's tolerance for falsely rejecting parity when it exists (Type I error is discussed in Section I.B below). Finally, the test designer must derive, from probability theory or known data, the probability distribution of the test statistic, describing the variability of performance under the null hypothesis.

14. Once these components are established, the test designer can determine (usually from a statistical table) a "critical value" against which to compare the computed value of the test statistic that is based on the actual results. If the test statistic is less than the critical value, it can be inferred that the ILEC's performance has "passed" the test of parity. If, however, the computed test statistic is

greater than the critical value, the ILEC's performance is judged to be not at parity, and the ILEC has "failed" the parity test for that measurement. The relationship between the performance distribution under the null hypothesis and the critical value is demonstrated graphically below.



A. Test Statistic: The Commission Should Use The Modified Z-Statistic Recommended By LCUG.

15. The modified z-statistic recommended by the Local Competition Users Group ("LCUG") is an excellent choice of test statistic in these circumstances. The "z-statistic" is a standard test statistic.[4] It is used to determine if the

[4] The formula for the z-statistic (also called the t-statistic), for the case where the observations are of measurements rather than proportions or rates, is

$$z = (\bar{Y} - \bar{X}) \Big/ \sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) S^2}$$

average results (or means) drawn from two separate performance samples (here the monthly ILEC performance data for itself and CLECs) have population means that are equal.[5] Thus, the standard z-statistic formula can determine whether, based on the reported results, the ILEC's average performance for itself and for CLECs is the same.

16. However, it is not enough to test for a difference in means alone. In order to obtain parity, CLECs are entitled to service from the ILEC that produces <u>both</u> the same mean performance and also the same variance in performance.[6] The z-statistic, in its standard form, is not

---

where $\bar{X}$ (resp. $\bar{Y}$) is the average of the ILEC (resp. CLEC) measurements, m (resp. n) is the number of these measurements, and S is a measure of the scale of variation of these measurements. The usual situation is that the statistical test is designed to detect a difference in the population means of the ILEC and CLEC measurements, assuming the population variances to be equal. In this case the standard choice for S^2 is

$$S^2 = S^2{}_{pooled} = \frac{(m-1)S^2{}_{ILEC} + (n-1)S^2{}_{CLEC}}{m+n-2}$$

[5] Similar statistics can also be used to detect differences in proportions and rates.

[6] The Commission also recognizes that it would be discriminatory if the ILEC has the same mean performance time for itself and CLECs but the variability of its performance for CLECs is greater (<u>see</u> <u>Notice, Appx. B, ¶ 4</u> ("variability of response times . . . may affect the competitiveness of a competing carrier but may not be reflected in a comparison of average response times")). For example, CLECs would be at a commercial disadvantage if ILEC retail customers could always rely on an installation period of 4 days while installation dates for CLECs ranged from 2-6 days or more.

designed to detect differences in variance between CLEC and ILEC performance.

17. In order to create a single test that can account for both of these factors, LCUG proposes a modification that will make the statistical test have the power to detect whether the ILEC's variance in its performance for CLECs is greater than the variance in its performance for itself. Specifically, LCUG proposes to use the ILEC variance, rather than the "pooled" variance, in calculating the z-statistic.[7] This proposal is based on well-supported statistical testing principles and combines the power of tests of means and tests of variance. Thus, if the test proposed by LCUG is used, there would be no need to develop a separate test of the equality of variances.[8]

18. Use of the LCUG modified z-statistic, rather than the more conventional form that uses a "pooled" variance, is appropriate here because the problem here is different from

[7] The LCUG proposal is to use $S^2=S^2_{ILEC}$. The resulting test statistic has the same distribution theory as the conventional one (using S^2_{pooled}) except for changing the "degrees of freedom" from m+n-2 to m-1. The effect of this change will be small if the parity hypothesis holds, since as the incumbent monopolist, the ILEC sample is likely to be much larger that the CLEC sample.

[8] See Notice, Appx. B, ¶ 4. It should also be noted that the use of separate tests for differences in averages and differences in variance would reduce the power of each separate test. Thus, it is preferable to use a single test that is sensitive to cases where both the mean and variance can increase.

that addressed in the standard texts. In the standard development, it is assumed that if the null hypothesis fails, it is only because the population means are different; the population variances are assumed to stay equal. This assumption is not appropriate here, because an increase in the CLEC variance would be a violation of parity, and the test should be able to detect it.[9]

19. As described above, the denominator of the formula for the z-statistic requires use of a figure for variance. Contrary to the suggestion of some ILECs,[10] the most appropriate variance to use in this case is the variance of the ILEC's performance for itself during the reporting period. This sample variance is the best available estimate of the variance of the ILEC process. Moreover, the entire purpose of the examination is to determine whether the ILEC is providing CLECs at least the same level of service as it provides to itself and its retail customers. Thus, for this

[9] Another standard form of the z-statistic is designed for the case where the two population variances may differ even under the null hypothesis. In this case one replaces $\left(\frac{1}{m}+\frac{1}{n}\right)S^2$ by $\frac{S^2_{ILEC}}{m}+\frac{S^2_{CLEC}}{n}$
This form of the statistic ii inappropriate here since under the parity hypothesis the two population variances are equal. Use of this form would reduce the probability of detecting violations of parity.

[10] I am informed that some BOCs have suggested that the variance used in the formula should be based solely on the variance experienced by the CLECs, and others have suggested the use of the pooled variance.

purpose, variance in the ILEC's performance is the standard against which the performance for CLECs should be measured.

B. The Error Probability Should Be Based On A One-Tailed Test With Type I Error At No More Than The 5% Level.

20. In determining an appropriate Type I error probability for the statistical test, it is important to recognize that any probability rate above 0% means that the statistical test will produce errors.[11] It is also important to understand that there are two distinct types of testing errors. "Type I" errors occur when a statistical test shows that two sets of results (here for the ILEC and CLEC) are inconsistent with the null hypothesis (i.e., are not in parity) when in fact the null hypothesis is true. "Type II" errors are the opposite. They occur when a statistical test indicates that the outcomes are in parity, but parity does not in fact exist. Both types of errors are possible.

21. There are two "tails" to Type I errors, but the. Notice (Appx. B, n.3) correctly notes that only one is pertinent here: errors relating to cases in which the ILEC's performance for CLECs is worse than its performance for itself. Under the Commission's rules, CLECs are entitled to performance that is "at least equal" to the performance the ILEC provides to itself. Those rules are not concerned with

[11] AT&T Statistical Ex Parte, p. B-1.

cases where, unintentionally, the ILEC provides CLECs with a level of performance that is better than the performance it provides to itself.[12] Thus, the Commission's rules themselves argue for a one-tailed test.

22. It should also be recognized that Type II errors are as real as Type I errors. Thus, there may be cases in which the ILEC is not in fact providing equal service to CLECs, but purely by chance the statistical test fails to reject the parity hypothesis. Thus, it is necessary to strike a balance between the two types of errors. If we choose to make the Type I error small, then the Type II error will be large; and conversely. AT&T proposes to set the Type I error at no more than the conventional level of 5%. This controls the frequency of false alarms to be at most 5% while making the probability of Type II errors small for violations that are of substantial size. Using a one-tailed test for Type I error at about the 5% level thus strikes a reasonable balance.[13]

[12] I am also informed that CLECs are not entitled to demand performance better than the ILEC provides to itself. Thus, there is no reason to believe that ILECs would intentionally provide their competitors with a higher grade of service than they provide to themselves and their retail customers.

[13] For general information supporting the 5% level, see AT&T Statistical Ex Parte, pp. B-1-B-2.

C. Probability Distribution Should Be Based On The T Distribution Or A Permutation Distribution Analysis.

23. For moderate or large sample sizes, it is appropriate to use the Student t (or "t") distribution to determine the critical value for the test. Use of this distribution, which is readily available in table form, is simple and straightforward and will produce statistically reliable results.

24. The published tables of critical values, using the t-distribution, are based on the assumption that the two populations (of ILEC and CLEC measurements) are exactly Normal. In practice, we will not have Normal distributions, and so these critical values are only approximations. There has been much debate as to the minimum sample sizes for which the tabulated values become acceptable approximations: numbers such as 10 or 30 have been suggested. But this must depend on the shape of the probability density function[14] of the populations, because there exist populations for which the approximation will never be adequate, even for very large sample sizes. In advance of reviewing the actual data, it is impossible to say for what size samples the tabulated values will be acceptable. Nevertheless, assuming that very large values of the observations do not occur and the populations have approximately symmetrical probability

[14] See the graph in ¶14 of these comments for an example of a probability density function.

density functions, I would guess that the tabulated values would be acceptable, provided that both the ILEC and CLEC samples have at least 10 members. Thus, the issue of sample size should not generally be a problem.

25. There is an alternative method for developing the probability distribution of the test statistic that can be used with smaller sample sizes.[15] Under this method, called the permutation distribution, the probability distribution is generated through use of the actual sample results, rather than a preexisting table. Given two samples, X's and Y's from ILEC and CLEC respectively, we combine these into one pool and then divide this into two sets X* and Y* in all possible ways. For each way, we calculate the corresponding z-score, say z*. This gives us a distribution of z* values, each of which is equally likely under the null hypothesis that the ILEC is treating customers impartially. Given the desired Type I error rate, we can read off the appropriate critical value and compare this with the observed value.

26. For example, if the data are

3 ILEC observations: X=1, X=2, X=4
2 CLEC observations: Y=3 and Y=5

[15] This method will provide reliable results for any sample size, but the use of the t-distribution and the associated table is simpler for all but very small sample sizes.

then the pooled set is (1,2,3,4,5) and there are 10 ways we can assign these five observations to the ILEC and CLEC samples. We get 10 values of z:

-2.74 -1.20 -0.60 -0.44 0.00 0.00 0.44 0.60 1.20 2.74

and the 5% critical value is 2.74. The actual observed value is 1.20, and so is judged to be not significant (i.e., we accept the null hypothesis).

27. This test procedure is valid irrespective of the form of the population distribution, since it depends only on the assumption that each possible permutation is equally likely under the null hypothesis.[16] The method can be used whenever the sample sizes are large enough to make the test statistic well defined, in the present case even for m=2, n=1.

28. The permutation distribution would be developed through the use of a computer program that would enumerate the samples necessary to generate the distribution. I wrote a program to perform this function in a commercially available program language called S Plus in one-half hour. Thus, I believe that a suitable program could be developed

[16] See, e.g., Cox and Hinkley, Theoretical Statistics (1974) (paperback edition Chapman and Hall, 1979), pp. 182-184; H. Scheffe, The Analysis of Variance (1959) John Wiley & Sons, Section 9.3; P. Good, Permutation Tests (1994) Springer.

promptly for use by the entire ILEC industry at minimal cost[17].

29. A resource issue relating to the use of the permutation distribution is the time needed to generate results. Unless the sample sizes are very small, the number of permutations to be generated is extremely large.[18] In order to deal with this problem, it would be reasonable to use a random sample of possible permutations to approximate the distribution. For example, if the number of possible permutations in a particular case exceeds 1000, the program could be designed to approximate the permutation probability distribution by randomly selecting 1000 permutations and constructing the distribution from those data. Because computers can perform calculations such as this with remarkable speed, the distribution for any measurement category could be ascertained within a few seconds.[19]

[17] The Cytel Software Corporation of 675 Massachusetts Avenue, Cambridge, MA, markets a product called StatXact which has the capability of performing permutation tests.

[18] If m=10, n=5, there are 3003 permutations; if m=20, n=10, there are over 30 million.

[19] The Notice (Appx. B, n.5) raises another interesting possibility for a statistical analysis of individual performance measurements, i.e., comparing the proportions of two samples that exceed some fixed value. AT&T is studying a variation of this concept, in which the fixed value is not specified in advance, but is determined from the ILEC sample itself. We use the upper 90% quantile of the ILEC sample to determine the level of service that the ILEC is providing for 90% of its customers and then measure what percentage of CLEC customers receive at least that level of service. The

D. ILECs' Compliance With Their Nondiscrimination Obligations Should Be Based On An Aggregate Assessment Of Parity.

30. One of the key concepts in the AT&T Statistical Ex Parte is that it is also appropriate to use statistical analysis to review the aggregate results of an ILEC's performance to determine whether it is in compliance with its nondiscrimination obligations. If we apply a large number, several hundred perhaps, of tests of individual performance measurement comparisons, each test having a Type I error rate of 5%, then we would expect, on average, about 5% of these tests to indicate non-compliance even when the ILEC is actually fully in compliance. Thus the fact that this many tests indicate non-compliance does not give conclusive evidence that the ILEC is not in compliance with its Section 251 nondiscrimination obligations. The number of tests that erroneously indicate non-parity will vary randomly about this average number. We need to derive some

"parity" hypothesis is rejected if the fraction of CLEC customers receiving that level of service is much smaller than the percentage of ILEC customers receiving such service. (For example, if the ILEC completes repairs on a specific service for 90% of its customers within 48 hours, parity is not achieved if the ILEC complete repairs for much less than 90% of CLEC customers within that amount of time.) This test procedure is non-parametric, i.e., it does not require any assumptions beyond the basic one that under the null hypothesis CLECs receive equal treatment to the ILEC. This methodology only applies, however, to the review of individual performance tests. It does not address the need to develop a method to review ILEC performance in the aggregate.

threshold number of failed parity tests such that if more than this number are observed to fail, then non-compliance can be deduced. This threshold number of tests must be determined in such a way as to control the probability of an overall, or aggregate, Type I error at 5%. Furthermore, I also recommend that any review of an ILEC's compliance with its nondiscrimination obligation should be based on two dimensions of statistical comparisons, both of which must be satisfied.[23] The two dimensions of statistical comparisons are

(a) the number of tests that fail in any monthly period must not be too large, and

(b) the number of tests that fail for three consecutive months must not be too large.

Here, "too large" must be determined by consideration of the total number of individual tests and the desired overall Type I error rate.

31. For the first dimension, we must determine how many of the individual measurements subjected to the above comparison tests need to demonstrate non-parity before an ILEC may be found to be in overall violation of its

[23] The AT&T Statistical Ex Parte suggested that a third dimension also be considered, namely imposing a bound on the number of individual tests that exhibit extreme violations. I now judge that imposing this additional constraint does not provide much additional power for detecting extreme violations, and in fact reduces the chance of detecting some more moderate violations.

statutory duty. Suppose we have made N individual tests, each having a 5% Type I error rate, and have found that K_1 of them indicate non-compliance. If K_1 is approximately .05 times N, we have no conclusive evidence of overall non-compliance. Under the assumption that the ILEC is in compliance, we can determine a number k_1 such that the probability that K_1 exceeds k_1 is 5%.[21]

32. The second dimension, <u>i.e.</u>, the number of measurements failing the test repeatedly, is necessary to assure that the ILEC failures are indeed random. Without this dimension, the ILEC might be able to "game" the process and produce repeatedly discriminatory results on measures that are critical to one or more competitors. Thus, for this dimension, we must determine how many individual measurements in an ILEC report may be allowed to fail the parity test in three successive months before finding that the ILEC has failed to provide parity.

33. Suppose we have made N individual tests for each of three months, each test having a Type I error of 5%. Let K_2 be the number of tests that have failed in all three months. The probability that any individual test fails in

[21] This computation assumes that under the null hypothesis, the number K_1 has a binomial distribution with exponent N, <u>i.e.</u>, it is as though we had tossed N coins, each with a probability of coming down "heads", and have counted how many "heads" appear. Then we claim non-compliance if K_1 exceeds k_1.

all three months, assuming that the ILEC is in compliance with its nondiscrimination obligation, is $(.05)^3$, or 1/8000. Thus the expected number that fail in all three months, assuming compliance, is N/8000. Given that the number of monthly tests will be well below 8000, noncompliance should be found if K_2 is not zero. In other words, the allowed number of three-time-failing tests is $k_2=0$.

34. If we apply both of these overall procedures simultaneously, the actual overall Type I error rate is a function of three things: the Type I error rates of the individual tests, which I call α_1, the number k_1 of allowed individual failures, and the number k_2 of allowed three-time failures. These three numbers can be determined so that the Type I error rate of the overall procedure is exactly 5% (or whatever other value is required). Details of this computation are given in Exhibit 1.

II. BellSouth's Proposed Methodology Is Unsuited To Measure Parity And Should Be Rejected.

35. The Notice (Appx. B, ¶ 7) also solicits comments on the methodology proposed by BellSouth, which is based on the use of statistical process control. This approach is not suitable to measure parity between ILECs and CLECs and should be rejected.

36. BellSouth has proposed three kinds of control charts. In the first, described in the Notice (Appx. B,

¶ 6), BellSouth maintains its own monthly results (presumably for each type of measurement) on a control chart. Three-sigma limits are established by reference to BellSouth's historical record. Then, each month, results for the CLEC are plotted on the same chart, and parity is claimed if these values do not fall outside the limits.

37. A second proposal appears in BellSouth's Tennessee Section 271 proceeding (see memo from David Laney to William Stacy, attached to the rebuttal testimony of William N. Stacy, TRA Docket 97-00309, Exhibit WNSPM-2). Here the proposal is to plot values of the variable DIFF=(CLEC value - ILEC value) on a control chart, with limits set at +/- 2.66 times the average moving range of size two.

38. A third proposal also appears in the same document from BellSouth's Tennessee Section 271 proceeding. Here it is proposed to compute z-scores, but using the process standard deviation in the denominator rather than the within-month ILEC sample variance as AT&T recommends. This process standard deviation is the average moving range (presumably of size two) divided by 1.128.

39. Each of these proposals has serious deficiencies, the most serious being that statistical process control is not designed to measure differences in parity. Rather, this technique is used to measure stability in performance. Stability of ILEC processes is of course an important

concept, because the overall reliability of the systems used to serve CLECs is essential to determining whether an ILEC has met its duties under Section 251 of the Act. However, it is irrelevant in determining whether an ILEC's performance for itself is at parity with the performance it provides to others, i.e., CLECs. The ILEC's performance could be stable, with parity not provided, or unstable with parity being provided. Stability and parity are distinctly different concepts.

40. Another shortcoming of each of the three BellSouth proposals is that no allowance is made for the fact that the number of observations that contribute to each average may change from month to month. This makes the use of moving ranges invalid measurements of variability. Also, the number of observations in the CLEC sample is very unlikely to equal the number in the ILEC sample. Thus the ILEC and CLEC averages will not have the same variances, even assuming parity, and so should not be compared to the same control limits, as the first proposal suggests.

41. If control limits for the quantity DIFF were to be set using the process variability of this quantity, as in the second and third proposals, some consistent violations of parity could completely avoid detection. Namely, if for any reason the CLEC measurements were consistently more variable than the ILEC measurements (which would imply that

many CLEC customers were getting poorer service), then this variability would be included in setting the control limits, and lack of parity would not be detected.

42. Further, use of separate control charts for each of the many types of measurement leaves open the question of how an overall judgement of compliance should be arrived at. BellSouth has not addressed this issue.

Conclusion

43. In summary, my testimony shows that AT&T's proposed methodology satisfies the Commission's desire to assure that reported differences in ILEC performance are statistically meaningful.

44. With respect to individual tests of ILEC performance, there are three key components in developing an appropriate statistical methodology. First, the modified z-statistic proposed by LCUG provides an appropriate test statistic to determine whether there are significant differences in the mean and the variance of an ILEC's performance for itself and for CLECs. Second, a one-tailed test with Type I error at about the 5% level strikes a fair balance between the need to account for both Type I and Type II errors. Third, the t-distribution provides a useful basis for calculating the critical value for individual tests of ILEC performance. Moreover, in those few cases

where the size of the ILEC sample is small, use of the permutation distribution will provide valid results.

45. It is also appropriate to aggregate the results of individual tests to determine whether the ILEC is in overall compliance with its duty to provide nondiscriminatory treatment to CLECs. This should be done through the use of a two-part analysis that sets limits on the number of individual tests that fail to demonstrate parity in any given month and on the number of individual tests that fail in three consecutive months. These limits can be determined in such a way that the overall Type I error is held at 5%.

46. Finally, the methodology suggested by BellSouth is not designed to measure parity of performance between two different populations. Thus, it should not be used to determine whether ILECs have met their legal duty to provide CLECs with parity service.

Colin L. Mallows

Colin L. Mallows

Sworn to before me this
29th day of May, 1998

Patricia A. Perhac
Notary Public

My Commission expires 4/8/2002

Exhibit 1

Statistical Definition of the Compliance Rule for ILEC Parity

The number k_1 of allowed individual violations, and the Type I error of each of the individual tests[22], α_1, are determined so that the probability of falsely claiming overall violation is controlled at a known level[23], which we call α.

Suppose we are aggregating N individual tests. Let K_1 be the number of these tests that indicate violations this month, and let K_2 be the number of tests that have shown violations in each of the past three months. Our proposed procedure is to claim overall violation if either (i) K_1 exceeds some number k_1, or (ii) K_2 exceeds zero. We show how k_1 and the type I error α_1 of each individual test can be determined so that the Type I error of the overall procedure is held at some desired level α.

To determine k_1 and α_1 when we know N, (the number of tests to be aggregated), and α, we proceed as follows. Throughout this calculation, we are assuming that the ILEC is fully in compliance, so that for each individual test the probability of (falsely) indicating non-parity is α_1.

a) Choose a tentative value for α_1. We start with $\alpha_1 = \alpha$. This value of α_1 will be adjusted (downwards) later.

b) Determine k_1 to be the largest number such that the probability that the overall procedure indicates violation[24] (is greater than α.

c) Decrease α_1 until the probability of overall violation using the value of k_1 that was determined in step b), is exactly α.

[22] Also referred to as the size of the individual test.

[23] Also referred to as the size of the overall aggregated test.

[24] This probability is: $1 - (1 - \alpha_1^3)^N * P(k, N, p)$ where P(,,) is the cumulative probability of the binomial distribution. That is, P(k, N, p) is the probability that the number of false parity test failures is <= k when the probability of an individual false parity test failure is p, and where $p = (\alpha_1 - \alpha_1^3)/(1 - \alpha_1^3)$.

The resulting value of α_1 (and the corresponding critical value on the z-score scale) is to be used in each of the individual tests. Then non-compliance is indicated if any series fails the test in three successive months, or if more than k_1 fail in any single month.

The following table provides an example of how k_1 is determined for the values N = 100 and α = 5%. As shown, the value of k_1 = 8 is the largest value of k that corresponds to a probability of no less than 5% of being exceeded. In this case, the probability of claiming an overall violation is 7.40%.

Table 1

Determination of k_1 for N=100, α= 5%

k	Prob{K_1>k, K_2>0} =1 - (1 - $\alpha_1^3)^N$ * P(k, N, p)
5	38.95%
6	24.17%
7	13.76%
8	7.40%
9	3.99%
10	2.36%

← select this k for k_1

The next step is to iteratively decrease α_1 and recompute the overall probability of violation, with k_1 held at 8, until we arrive at a value for α_1 for which this probability is .05. In this case, that value of α_1 is .04601.

Now we can use the t-tables (or permutation distribution calculations) to determine the appropriate critical values for each individual test. The following Table 2 provides k_1, α_1, and critical values (assuming large sample sizes for each test) for α = .05 and a number of values of N.

Table 2

Determination of k_1 and α_1 for a range of N
where k_1 satisfies $1 - (1 - \alpha_1')^N * P(k_1, N, p) = .05$

N	k_1	k_1 as a % of N	α_1	Critical Value (c)
70	6	8.57%	.0465	1.6803
80	[illegible]	7.50%	.0408	1.7411
90	7	7.78%	.0437	1.7096
100	8	8.00%	.0460	1.6849
120	9	7.50%	.0442	1.7038
140	10	7.14%	.0430	1.7170
160	12	7.50%	.0462	1.6825
180	13	7.22%	.0452	1.6937
200	14	7.00%	.0443	1.7026
250	17	6.80%	.0441	1.7046
300	20	6.67%	.0440	1.7060
400	26	6.50%	.0437	1.7095
500	32	6.40%	.0431	1.7155
600	38	6.33%	.0423	1.7247
700	44	6.29%	.0412	1.7374
800	49	6.13%	.0397	1.7543
900	55	6.11%	.0384	1.7696
1000	60	6.00%	.0371	1.7851

D

KPMG Consulting has been asked to provide an assessment of the *Indiana & Ohio CLECs Statistics Statement of Position* (distributed to the Ohio and Indiana Collaborative members) and the *Petition for Rehearing or Reopening* of the Wisconsin statistical order (collectively the "*OH/IN/WI Request*"). This assessment is presented below.

The *OH/IN/WI Request*:

1. Proposes commission-mandated implementation of suggestions that were not agreed upon during the collaborative development process;

2. Does not reflect KPMG Consulting's view of certain elements of Appendix C and the Statistical White Paper in Wisconsin and the current proposed Appendix C in Ohio and Indiana;

3. Significantly alters the baseline project schedule;

4. Does not produce any significant improvement in the Evaluations.

The *OH/IN/WI Request* proposes commission-mandated implementation of suggestions that were not agreed upon during the collaborative development process.

The respective state regulatory commission staffs, Ameritech and the CLECs have collaborated for the past several months to develop and refine the statistical approach described in the proposed Appendix C distributed to Collaborative members in Ohio (March 1, 2001) and Indiana (March 2, 2001), respectively, and Appendix C and the Statistical White Paper in Wisconsin[1]. KPMG Consulting participated in these collaborative sessions consistent with its role as an independent third-party test manager.

During the development of Appendix C in these states, Collaborative members discussed suggestions of dual hypothesis testing, exchanging the null and alternative hypotheses and the use of full sample sizes at all disaggregations or for select CLEC-specified measures before the current version of Appendix C was produced. The current version of Appendix C was refined to account for Collaborative participants' concerns with some elements of the statistical approach described in the original version of Appendix C. Items were addressed through updates to the original proposed version of Appendix C and by development of the Statistical White Paper in Wisconsin, which has been integrated into the current proposed Appendix C in Ohio and Indiana. Essentially, the *OH/IN/WI Request* appears to abandon Collaborative-driven enhancements to the statistical methodology and advocates that state regulatory

[1] The proposed Appendix C in Ohio and Indiana is a combination of the accepted Wisconsin Appendix C and the Statistical White Paper. In total, each of these documents or sets of documents describe exactly the same statistical approach. For the sake of simplicity, references to the accepted Appendix C in Wisconsin throughout the remainder of this assessment implicitly acknowledge the inclusion of the Wisconsin Statistical White Paper as part of Appendix C.

commissions order the implementation of suggestions that were not acceptable to all Collaborative parties.

The *OH/IN/WI Request* does not reflect KPMG Consulting's view of certain elements of Appendix C.

KPMG Consulting disagrees with the following points in the *OH/IN/WI Request*:

1) "The test's statistical methodology employs incorrect hypotheses;"

2) "The statistical methodology does not balance error potential."

Correctness of the Hypotheses. Appendix C relies on a Null Hypothesis of parity for metrics that involve parity because parity is the only clear and easily definable hypothesis. Additionally, in cases of identical, or supposedly identical, processes (such as those involving parity measures), the logical methodology is to attempt to disprove the Null Hypothesis of parity. As a result, these Evaluations are designed to focus on testing for credible evidence of the absence of parity, rather than to "prove" the existence of parity. Hence, the *OH/IN/WI Request* incorrectly interprets the application of the scientific method in this case.

Balancing of Error Potential. The *OH/IN/WI Request* asserts an imbalance between the error rates set for Type I (false failure) and Type II (false pass). The *OH/IN/WI Request* claims the concept of a Type II error rate of 50% for disaggregations is "no more scientific than a coin flip." This assertion is tantamount to claiming that all statistical tests are unscientific. Despite implicit claims to the contrary in the *OH/IN/WI Request,* every standard statistical test of means has some level of disparity where the error rate is 50%.

Appendix C sets sample sizes to equate Type I and Type II error rates for each metric at the aggregated level, for a level of disparity agreed upon at one point by Collaborative participants. The Type I and Type II error levels are set at 5% for specific "aggregated" metrics in these states. The Type II error rate is set at 50% for this same level of disparity for "disaggregated" metrics. Appendix C details the reasons for the error rates that correspond to the particular level of disparity specified.

Despite the rationale provided in Appendix C, Type I and II error levels continue to be the subject of great misunderstanding in this instance. Logically, if product disaggregations exist within an aggregated metric, error rates for these disaggregations need to be higher than for the aggregated metric itself because the sample size is necessarily smaller. Type I and Type II errors could be equalized for all disaggregated metrics. However, such activity would result in a much smaller Type II error rate than Type I error rate for aggregated metrics. Furthermore, this suggestion was raised and rejected earlier in the collaborative process because it significantly increases the size of the test beds, and therefore significantly increases the baseline project schedules.

The *OH/IN/WI Request* significantly alters the baseline project schedule.

KPMG Consulting's baseline project schedules for Wisconsin, Ohio and Indiana (currently under development) are predicated on the current version of Appendix C. Exchanging the null and alternative hypotheses (as they are currently proposed) or utilizing dual hypotheses will necessitate a substantial outward movement of the baseline project schedules estimated at four to six months.

A null hypothesis that relies on the principle of equality (which is the case with the current version of Appendix C) is elegantly simple and robust. Equality is, after all, equality. There is no reason to establish the "level" at which equality exists for each test. Thus, Appendix C, as accepted in Wisconsin and as proposed in Ohio and Indiana, does not require any additional development before testing can begin.

Exchanging the null and alternative hypotheses or engaging in dual hypothesis testing is deceptively complicated. A null hypothesis predicated on disparity is much more problematic because the level of disparity must be established for each and every test. This determination is necessary because the effect of disparity, even measured in standard units of variability (such as standard deviation), is likely to differ by product. Additionally, Commission and Collaborative guidance will be vital in establishing the appropriate levels of disparity for the Ameritech products. KPMG Consulting could recommend levels for each metric, but such proposals would involve an extended analysis of Ohio, Indiana and Wisconsin data in advance of the test. Alternatively, it has been suggested previously that one level of disparity as a function of standard deviation should be employed for all tests. While this is convenient, the suggestion is unjustified.

Based on Master Test Plan and statistical methodology development cycles in Indiana, Ohio and Wisconsin, the time required to perform data analysis and the likelihood of debate regarding the appropriate levels of disparity, KPMG Consulting estimates that the process of developing the appropriate levels of disparity could extend the baseline project schedules from four to six months (since statistical methodology is a critical path item). Additional unforeseen delays during testing, due to inexperience with this methodology, are also possible.

Likewise, dual hypothesis testing will be subject to the same estimated four to six month schedule delays given the need to establish the levels of disparity for the second hypothesis.

The *OH/IN/WI Request* does not produce any significant improvement in the Evaluations.

The *OH/IN/WI Request* concludes that dual hypothesis testing is necessary, if recommendations for exchanging the hypotheses and equalizing the risk of Type I and Type II errors are not granted. As described previously, the concern regarding balancing Type I and II errors for disaggregated measures is based on a statistical interpretation with which KPMG Consulting does not agree and whose only redress was deemed untenable for these Evaluations. Likewise, the exchange of hypotheses or use of dual hypotheses is both statistically unnecessary and burdensome in the specific instance of these OSS evaluations. However, the request for additional analysis and reporting of those analyses is well-grounded in certain instances.

To this end, the current version of Appendix C prescribes additional analysis and reporting when sample sizes are not met or when other inconsistencies in the data or test arise. This additional analysis and reporting was added to Appendix C during the Collaborative's development process to address specifically this request for such activity, when necessary.

In summary, the statistical approach described in the current version of Appendix C has evolved significantly as a result of KPMG Consulting's testing experiences and the multiple Collaborative sessions. Appendix C illustrates a statistical approach that is sufficiently robust to yield meaningful results without adding time and cost to the Ameritech OSS Evaluations in Indiana, Ohio and Wisconsin.